Project 2

Clara Torslov (ct32699)

This is the dataset we will be working with:

```
members <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2
```

```
library(cowplot)
library(colorspace)
library(MASS)
library(tidyverse)
library(ggridges)
library(ggforce)
```

Part 1

Question: Looking only at expeditions to Mt.Everest since 1960, how do deaths in each season break down by the seven most common causes?

Introduction: We are working with the data set, `members`, a historical data set about Himalayan expeditions, containing records for all individuals who participated in expeditions from 1905 through Spring 2019 to more than 465 significant peaks in Nepal. The data is taken from the Himalayan Database, which is a compilation of records for all expeditions that have climbed in the Nepal Himalaya. More information about the dataset can be found at https://github.com/rfordatascience/tidytuesday/blob/master/data/2020/2020-09-22/readme.md and https://www.himalayandatabase.com/. Each row in the data set corresponds to an individual who participated in an expedition and the columns represent different information about that individual.

Throughout part 1 we consider the variables `death_cause`, `season`, `peak_name` and `year`. The latter variable tells which year the expedition took place and `season` tells which season the expedition took place (Spring, Summer, Autumn, Winter). `peak_name` contains information about the name of the mountain, fx the value `Everest` tells us the expedition took place on Mount Everest. `death_cause` tells the primary cause in case of death. If the individual did not die on the expedition the value will be `NA`.

Approach: To answer the first part of this question we will perform data wrangling to create an informative subset of the data set. We wish to use `filter()` to only get the expeditions to Mt.Everest since 1960. As we wish to examine death causes we wish to remove missing values as they indicate that the individual did not die. We further wish to only consider the seven most common causes of death separately for each season and collect the remaining death causes in a `Other` value. We therefore wish to use `mutate()` and `fct_lump_n()` before we count each death cause for each season (using `count(death_cause, season)`). We present the result in a summary table using `pivot_wider()` and make sure to replace any `NA` values with `0`.

Then, to compare the proportions between the different death causes, we wish to do a pie chart. This is great as we wish to examine how the individual death causes compare as a fraction of the whole, and works well for smaller data sets, like ours.

Analysis: We perform data wrangling to create a summary table.

```
members %>%
  filter(died == TRUE &
           year >= 1960 &
           peak_name == "Everest") %>%
  mutate(
```

```
    death_cause = fct_lump_n(fct_infreq(death_cause), 7, other_level = "Other"),
    season = fct_relevel(season, "Spring", "Autumn", "Winter","Summer")
        ) %>%
  count(death_cause, season) %>%
  pivot_wider(names_from = season, values_from = n) %>%
  mutate_all(~replace(., is.na(.), 0))
```

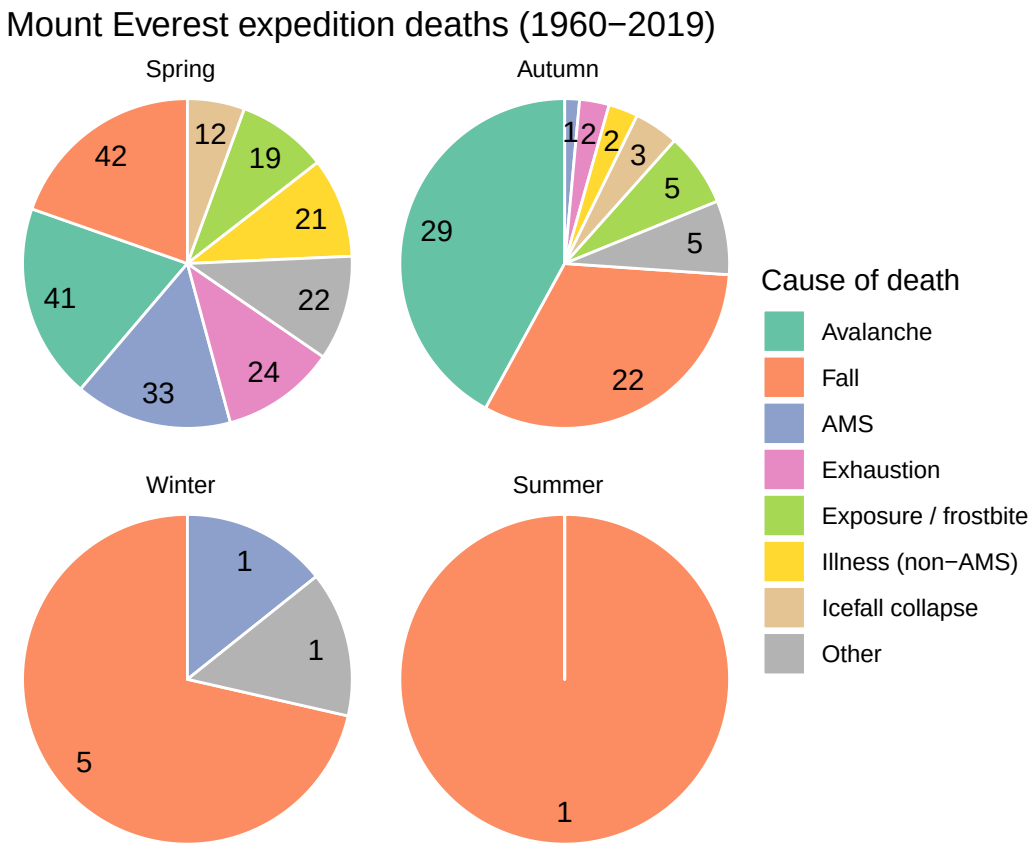
```
## Warning in `[<-.factor`(`*tmp*`, list, value = 0): invalid factor level, NA
## generated
```

```
## # A tibble: 8 x 5
##   death_cause          Spring Autumn Winter Summer
##   <fct>                 <dbl>  <dbl>  <dbl>  <dbl>
## 1 Avalanche                41     29      0      0
## 2 Fall                     42     22      5      1
## 3 AMS                      33      1      1      0
## 4 Exhaustion               24      2      0      0
## 5 Exposure / frostbite     19      5      0      0
## 6 Illness (non-AMS)        21      2      0      0
## 7 Icefall collapse         12      3      0      0
## 8 Other                    22      5      1      0
```

We plot a pie chart to examine the proportions of each death cause for each season.

```
members %>%
  filter(!is.na(death_cause) &
           year >= 1960 &
           peak_name == "Everest") %>%
  mutate(
    death_cause = fct_lump_n(fct_infreq(death_cause), 7, other_level = "Other"),
    season = fct_relevel(season, "Spring", "Autumn", "Winter","Summer")
        ) %>%
  count(death_cause, season) %>%
  arrange(n) %>%
  group_by(season) %>%
  mutate(
  end_angle = 2*pi*cumsum(n)/sum(n),
  start_angle = lag(end_angle, default = 0),
  mid_angle = 0.5*(start_angle + end_angle),
  hjust = ifelse(mid_angle > pi, 1, 0),
  vjust = ifelse(mid_angle < pi/2 | mid_angle > 3*pi/2, 0, 1)
  ) %>%
  ggplot() +
  aes(x0 = 0, y0 = 0, r0 = 0, r = 1,
      start = start_angle, end = end_angle,
    fill = death_cause) +
  geom_arc_bar(color = "white") +
  coord_fixed() +
  geom_text(
    aes(
      x = 0.8 * sin(mid_angle),
      y = 0.8 * cos(mid_angle),
      label = n),
    size = 4
    ) +
```

```
  facet_wrap(vars(season)) +
  theme_void() +
  theme(strip.text.x=element_text(margin=margin(b=1, t = 5, r = 5))) +
  ggtitle("Mount Everest expedition deaths (1960-2019)") +
  guides(fill = guide_legend("Cause of death")) +
  scale_fill_brewer(palette = "Set2")
```



Discussion: From the summary table we see that `Avalanche` is the most common cause of death and that the highest frequencies of deaths occur during `Spring`. Looking at the pie charts we see that during Autumn we see that the highest proportion of deaths on Mt. Everest since 1960 is due to an Avalanche, closely followed by the second largest proportion, `Fall`. Besides these two large proportions we see 6 smaller fractions of various death causes. During Spring we see a different kind of spread among the death causes. During this season the different death causes are in general more evenly proportioned. With fall being the largest proportion, almost equal to deaths by an Avalanche. The smallest proportion of deaths during Spring is due to icefall collapse. We see a smaller number of deaths in general during Winter and Summer. During Summer we only see a single death due to falling. During Winter there has been a total of 7 deaths, with falling being the most dominant cause of death.

Part 2

Question: Which peaks has been the most successful to climb? Does the distribution of success to these peaks change for different ages and expedition group sizes?

Introduction: We are still working with the `members` data set. In this problem we consider the variables `peak_name`, `success`, `expedition_id` and `age`. `peak_name` contains information about the name of the mountain. `success` is a logical variable indicating whether the individual was successful in summitting the peak. `expedition_id` is the unique identifier for each expedition and indicates which expedition the individual took part in. `age` states the age of the individual.

Approach: To answer the question posed in the second part, we first wish to create a summary table with the top 25 most successful peaks to climb, showing the peak and the success of summitting this peak. With data manipulation techniques we wish to create a new variable, which denotes the rate of success for each peak, based upon the `success` variable and another new variable, which denotes the expedition size, based upon the `expedition_id` variable. We only consider peaks which has been attempted to summit a decent amount of times (50 attempts). We further wish to examine expeditions to these peaks with a ridgeline plot. We wish to do so as they are often useful when visualizing and comparing a larger number of distributions across several groups and they will accurately represent bimodal data.

Analysis:

We perform data wrangling to create a summary table, and a subset, with the most successful peaks.

```
(successPeaks <- members %>%
  filter(!is.na(peak_name)) %>%
  count(peak_name,success) %>%
  pivot_wider(names_from = success, values_from = n) %>%
  mutate_all(~replace(., is.na(.), 0)) %>%
  rename("NoSuccess" = "FALSE",
         "Success" = "TRUE") %>%
  group_by(peak_name) %>%
  mutate(
    SuccessRate = Success / (Success + NoSuccess)
  ) %>%
  filter(NoSuccess + Success > 50) %>%
  arrange(desc(SuccessRate)) %>%
    head(n = 25)
  )
```

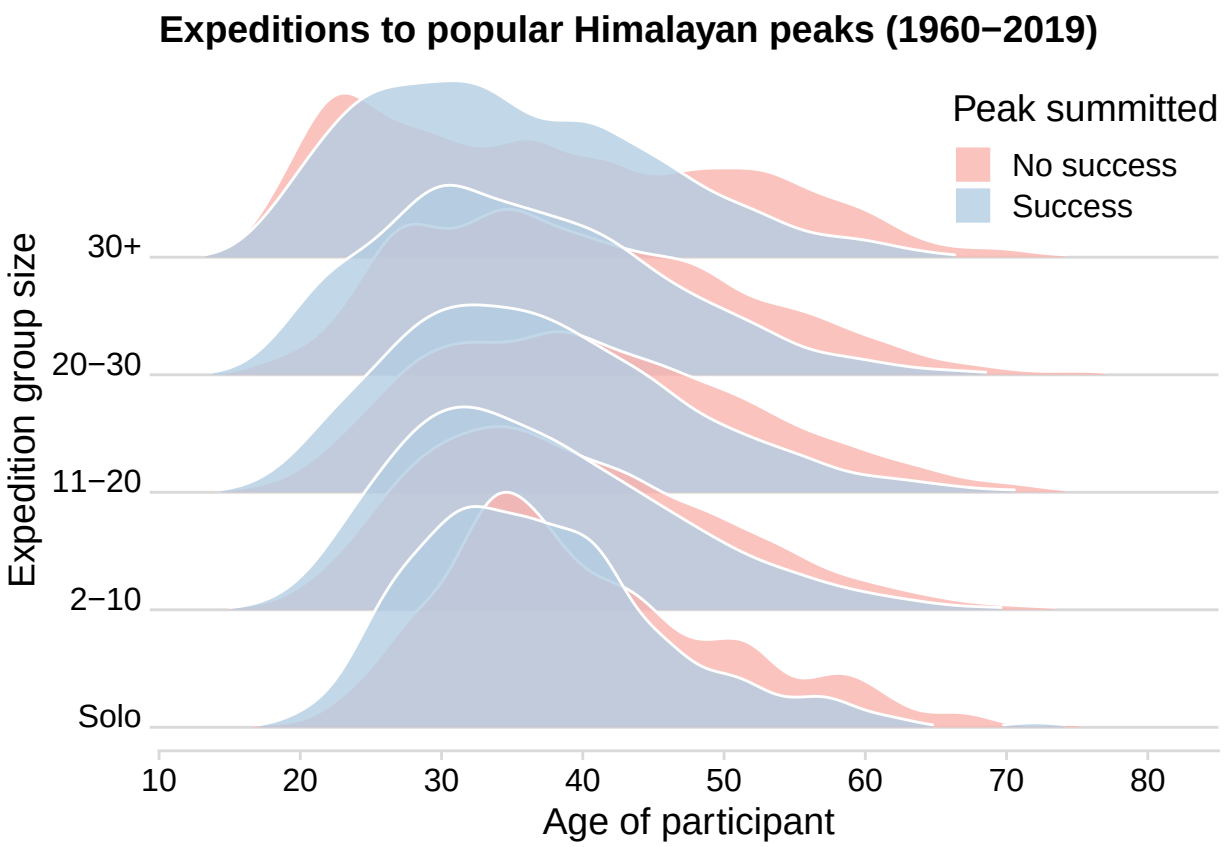
```
## # A tibble: 25 x 4
## # Groups:   peak_name [25]
##    peak_name             NoSuccess Success SuccessRate
##    <chr>                     <dbl>   <dbl>       <dbl>
##  1 Saribung                     31      79       0.718
##  2 Dhampus                      47      81       0.633
##  3 Tengkoma                     21      31       0.596
##  4 Urkema                       42      52       0.553
##  5 Ama Dablam                 4023    4383       0.521
##  6 Arniko Chuli                 26      28       0.519
##  7 Lamjung Himal                42      41       0.494
##  8 Everest                   11777   10036       0.460
##  9 Kangchenjunga South          34      27       0.443
## 10 Kangchenjunga Central        32      25       0.439
## # ... with 15 more rows
```

We `left_join()` above subset with the `members` data set to get information for the top 25 most successful peaks only. We then plot the distribution of age vs expedition size between success and non-success, using ridgeline plots.

```
successPeaks %>%
  left_join(members) %>%
  filter(!is.na(highpoint_metres)) %>%
  group_by(expedition_id) %>%
  mutate(
    expedition_size = n(),
    expedition_size_group = case_when(
      expedition_size == 1 ~ "Solo",
```

```
      expedition_size > 1 & expedition_size <= 10  ~ "2-10",
      expedition_size > 10 & expedition_size <= 20   ~ "11-20",
      expedition_size > 20 & expedition_size <= 30   ~ "20-30",
      expedition_size > 30 ~ "30+",
    ),
    expedition_size_group = factor(expedition_size_group,
                                   levels = c("Solo", "2-10", "11-20","20-30","30+")
                                   ),
    highpoint_metres_group = case_when(
      highpoint_metres > 8000 ~ "> 8000",
      highpoint_metres <= 8000 & highpoint_metres > 7000 ~ "7001-8000",
      highpoint_metres <= 7000 & highpoint_metres >= 6000 ~ "6000-7000",
      highpoint_metres < 6000 ~ "< 6000"
    ),
    highpoint_metres_group = factor(highpoint_metres_group,
                                    levels = c("< 6000", "6000-7000", "7001-8000","> 8000")
                                    )
  )  %>%
  mutate(success = ifelse(success == TRUE, 'Success', 'No success')) %>%
  filter(!is.na(age)) %>%
  ggplot() +
  aes(x = age,
      y = expedition_size_group,
      fill = success) +
  geom_density_ridges(color = "white",
                      rel_min_height = 0.01,
                      alpha = 0.8,
                      bandwidth = 2,
                      scale = 2) +
  scale_y_discrete(name = "Expedition group size",
                   expand = expansion(add = c(0.2, 1.6))
                   ) +
  scale_x_continuous(name = "Age of participant",
                     breaks = seq(10,80, by = 10),
                     limits = c(10,85),
                     expand = c(0, 0)
                     ) +
  theme_minimal_hgrid() +
  ggtitle("Expeditions to popular Himalayan peaks (1960-2019)") +
  labs(fill = "Peak summitted" ) +
  theme(
    axis.text.y = element_text(vjust = 0),
    axis.title = element_text(hjust = 0.5),
    legend.position = c(1, 0.77),
    legend.justification = c(1, 0),
    plot.title = element_text(size = 14, hjust = 0)
  ) +
  scale_fill_brewer(palette = "Pastel1")
```

```
## Joining, by = "peak_name"
```



Discussion: From the summary table we see that Saribung is the most successful peak to climb with a success rate of $71,8\%$ and that 6 peaks have a success rate above 50%. Mount Everest is the 8th most successful peak to climb. Further examining what entails success to these peaks we see from the ridgeline plot that age has a clear effect. We see this as the distribution of `No success` is clearly shifted slightly to the right compared to `Success`, for all expedition group sizes, indicating more successful younger participants. Considering the right tail of the `Success` distribution we see that the medium expedition group size (`11-20`) stretches furthest to the right, meaning if an individual is up there age-wise it would be advisable to take part in an expedition of this size. Still considering the `Success` distributions we see that, across all expedition group sizes, the highest probability of success is around 30 years (of age). The `Success` distribution shape differs across expedition group sizes. The peak of the mode of the distribution shifts slightly to the left as the expedition group size increases, meaning as group size increase, there is a higher probability of success for younger individuals. The only exception here is the largest group size `30+`, where individuals below the age of 25 (approximately) have a larger probability of no success than success. In general we see that most individuals who have attempted to climb these peaks are in the range 25 to 45 years, with most individuals being in their 30's.